WEB掲示板の書き込みを利用したラーメンに関する地域性の量的評価

WEB掲示板の書き込みを利用したラーメンに関する地域性の量的評価

WEB掲示板の書き込みを利用したラーメンに関する地域性の量的評価

村瀬浩章[1)]、澤井政宏[2)]、岡田吉史[3)]、長島知正[1,3)]
1）室蘭工業大学　情報工学科
2）室蘭工業大学　工学研究科生産情報システム工学専攻
3）室蘭工業大学　サテライト・ベンチャー・ビジネス・ラボラトリー

1. はじめに

各地域の特産物や地域限定の食品、味付けの違いなど、食の嗜好には地域性が存在する。食の地域性は、食品メーカが消費者の嗜好に合うように商品の開発や改良を行う際の重要な情報となっている。また、そのような情報はオンラインショッピングの商品推奨機能（Amazon.com など）等において、個人の感性に見合う商品を提供するための有用な情報としても利用されうる。

近年、インターネット上の多くの WEB 掲示板において、地域ごとの様々な食品に関する議論が盛んに行われるようになった。WEB 掲示板の書き込みでは、投稿者は、1）自由記入で感想を書き込めること、2）評価に参加する意思を持って書き込むことなどから、投稿者の食品に対する評価が直接的に反映されていると考えられる。また、各地域の WEB 掲示板には、各々の地域の食品をよく特徴付けるキーワードが頻繁に出現している可能性が高く、食品の嗜好に関する地域性を定量する上で有用な情報源になると考えられる。

本研究では、WEB 掲示板の書き込みの単語統計に基づいて、各地域を特徴付けるキーワードを抽出し、これによって食品の嗜好に関する地域性を評価する方法を提案する。以下、本稿では、地域によって味付けや具材の違いが比較的明瞭であり、広く国民に親しまれているラーメンを例として、味付けや特徴的な具材などに関する地域性の評価を行う。

2. 単語の統計解析

本研究では、食品の嗜好に関する地域性を量的に評価するために、WEB 掲示板の書き込みに出現する単語の統計的性質に基づいて、単語の重み付けを行った。各地域において、重要度が高いと判断された単語（重みが大きい単語）は、その地域のラーメンの地域性を良く表すことが期待される。我々は以下の 2 つの手法を用いて単語の重み付けを行った。

2.1 TF-IDF 法

TF-IDF 法[1]とは、(1)対象となる文書における単語の頻度（TF）と、(2)その単語が他の文書でどれだけ出現しないか（IDF）によって単語の重み付けを行う手法である。具体的に単語 t の重要度は以下のように計算される。

$$\text{重要度} = tf(t) * \log \frac{N}{df(t)}$$

ここで、tf(t)は単語 t の頻度、N は比較文書数、df(t)は N の中で単語 t が含まれている文書数である。

2.2 CWD 法

CWD 法[2]とは、単語 t を含む文に出現する他の単語（以下共出語）に着目して、共出語の種類が少ない単語ほど高い重要度を持つように重み付けを行う手法である。具体的に単語 t の重要度は以下のように計算される。

$$\text{重要度} = a - b * \text{共出語種数} / \text{単語} t \text{の頻度}$$

ここで a と b はパラメータであり実数値を取る。a が大きくなれば共出語種数が単語の重要度に及ぼす影響が小さくなり、b が大きくなると、共出語種数が単語の重要度に及ぼす影響が大きくなる。本研究では、a=1、b=1 とした。

3. 手法

以下、各地域の WEB 掲示板から、ご当地ラーメンを特徴付ける単語を抽出し、評価する方法について説明する。まず、WEB 掲示板の書き込みの本文のみを抽出するため、「名前」、「投稿日」などのヘッダ部分を除去した。

次に、得られた書き込みの本文に対して形態素解析を行った。形態素解析とは、与えられた文を品詞ごとに区切る手法である。ラーメンの味付けや具材に関する単語は「名詞」に含まれるため、それ以外の品詞を持つ単語を除去した。また、頻度 2 以下の単語は重要ではないと判断して除去した。次に、TF-IDF 法と CWD 法を用いて単語の重み付けを行い、それぞれの値に対して平均0、分散1となるように正規化を行った。最終的に、抽出された単語の重要度にしたがって「味」や「具」などのカテゴリ別にレーダーチャートを作成し、ラーメンの地域性を視覚的に表現する。

4. 実験

4.1 実験1

TF-IDF 法と CWD 法の 2 つの重み付け手法について、各地域の特徴を表す単語が適切に抽出されているかどうかを評価する。ここでは、正解キーワードとして各地域を特徴付ける単語を手作業で抜き出し（以下特徴語）、それらを本手法により抽出された単語と比較することにより評価を行う。

4.1.1 実験手順

実験データには、札幌、旭川、博多のそれぞれのラーメンに関する WEB 掲示板を使用した。まず、3 章で記述した手法によって単語毎の TF-IDF 値と CWD 値を算出し、それぞれの値を正規化した。

次に、上記のご当地ラーメンを紹介している 3 件の WEB

表1:各地域の特徴語

札幌	旭川	博多
ミソ	ショウユ	トンコツ
ラード	チャーシュー	ネギ
チャーシュー	サカナ	ベニショウガ
モヤシ	アブラ	カエダマ
ニンニク	ネギ	タカナ(高菜)
ネギ	メンマ	ストレート
トンコツ	トンコツ	
メンマ		
アブラ		

サイトに共通して出現する単語（名詞）を手作業により選出し、それらを各地域のラーメンを特徴付ける特徴語とした。表.1 に各地域から選出した特徴語を示す。続いて、この特徴語と、TF-IDF 値ならびに CWD 値の高い順にそれぞれ上位 100 位の単語（以下上位単語）を用いて再現率ー適合率を算出した。再現率、適合率は共に情報検索分野の標準的な性能評価基準である。再現率はその地域のすべての特徴語に対する上位単語に現れた特徴語の割合であり、適合率は上位単語に現れた特徴語の割合である。我々はこの再現率、適合率を比較して、各重み付け手法の性能を評価する。

4.1.2 実験結果

TF-IDF法とCWD法の再現率ー適合率を表.2、表.3に示す。表.2、表.3 に示されるとおり、TF-IDF 法の性能が CWD 法の性能を上回っている。このことから本研究において単語の重み付け手法としては、TF-IDF 法のほうが適切であるといえる。

4.1.3 考察

TF-IDF 法に比較して CWD 法の結果が劣っているのは、WEB 掲示板の書き込みには文章として不完全な文が多数含まれていたためと考えられる。本来、CWD 法で用いるデータは正しい構文で書かれた文章が前提になっている。また、WEB 掲示板の書き込みには単一の単語のみからなる文や、アスキーアートと呼ばれるアスキー文字から作られる絵文字がある。そのため CWD 法では正確に共出語が抽出できず、単語の重み付けが適切に行われなかったことも TF-IDF 法の結果を下回った原因であると思われる。

4.2 実験 2

以下、TF-IDF 法に基づく結果に限定して議論をする。

4.2.1 実験手順

4.1 で得られた TF-IDF 値を正規化した値を用いて、レーダーチャートを作成する。このレーダーチャートにより、ラーメンに関する地域性を視覚的に評価することができる。

4.2.2 実験結果

一例として札幌のラーメンの「味」（味噌、塩、醤油）について作成したレーダーチャートを図.1 に示す。各軸は正規化された TF-IDF 値を表す。

4.2.3 考察

図.1 が示すとおり札幌では「味噌」の重要度が最も高くなった。一般的に、札幌は味噌ラーメンが有名だと言われているため、これは妥当な結果であると考えられる。また、「醤油」、「塩」ともに「味噌」ほどではないが高い重要度を示している。札幌ではラーメンを観光の目玉の一つとしているため、函館ラーメン（塩）、旭川ラーメン（醤油）など様々な地方のラーメン店が出店している。そのため、それらが相まって 3 つの味がバランスよく着目されていると考えられる。このように、本手法において大雑把ではあるが、WEB 掲示板の投稿者による札幌ラーメンの評価を量的に表現できたのではないかと考えられる。

表.2：TF-IDF 法における再現率ー適合率

	札幌	旭川	博多
再現率	78%	57%	83%
適合率	7%	4%	5%

表3:CWD 法における再現率ー適合率

	札幌	旭川	博多
再現率	44%	43%	67%
適合率	4%	3%	4%

図1:札幌における「味」のレーダーチャート

5. まとめ

本稿では、各地域の食品に関する特色、すなわち地域性を把握することを目的として、WEB 掲示板の書き込みの単語統計に基づいて地域性を量的に評価することを試みた。実験により、本手法において WEB 掲示板における単語の重み付けには TF-IDF 法のほうが CWD 法よりも優れていることがわかった。さらに、レーダーチャートが示すとおり、大雑把にではあるがご当地ラーメンの地域性の量的評価を行うことができた。

今後はさらに食の地域性を量的評価するのにふさわしいレーダーチャートの軸となる単語を統計学的な分析などにより選定していくことが必要となる。

参考文献

[1] Salton, G. and Buckley, C.: "Term-weighting approaches in automatic text retrieval," Information Processing and Management, 24, pp.513-523, 1988.

[2]FUJITSULIMITED:http://venus.netlaboratory.com/salon/chiteki/jfs/index.html/ (1999-2005).